

ACE RESEARCH NOTE

The Approval Button Is Not Authorization

ACE-RN-2026-013 · Version 1.0

TEMPLATE PREVIEW

AUTHOR

Chris Ma

PUBLISHED

2026-08-18

VERSION

1.0

RESEARCH SCOPE

Human approval integrity · Tool & data authorization · Decision evidence completeness

PART I

Problem Definition

EXTERNAL RESEARCH

NIST and EU guidance require effective human oversight, defined roles, understanding of limitations, and ability to override or stop; OWASP separately requires downstream authorization and high-impact approval [R1] [R2] [R3].

ACE OBSERVATION

ACE Observation: a feature name, successful request, or user-interface state is not sufficient unless the tested system can reproduce the control behavior and its evidence.

A click records an interface event but does not prove that the reviewer was authorized, saw the final action and evidence, understood the scope, or that the backend enforced the approved object [R1] [R2].

This note treats the approval button is not authorization as a bounded control question. It does not infer universal product behavior from a paper, standard, interface screenshot, or single test. A demonstrated result applies only to the cited configuration; an authoritative source defines a requirement or design direction but does not certify an implementation. ACE therefore asks whether the tested system can produce the required behavior and evidence under declared versions, policies, topology, identities, and failure conditions.

The operational distinction is between a claim and a control that can be re-performed. The positive control is: An authorized reviewer signs action digest A1 after seeing its evidence and A1 executes once before expiry. The adversarial condition is: Parameters change after approval, the reviewer lacks authority, the approval is replayed, or the backend never checks it. A useful result must show both that authorized work remains possible and that the prohibited path is stopped before an irreversible side effect. Missing fields are classified as insufficient evidence, not silently converted into a pass.

Real-world impact

- A ceremonial click can transfer perceived responsibility without controlling the machine action.
- Approval fatigue can normalize confirmation of changes the reviewer cannot meaningfully inspect.
- A control claim without versioned evidence can mislead procurement, audit, incident response, and system owners.

- Failing closed without a positive control can conceal a denial-of-service design rather than demonstrate trustworthy behavior.

PART II

Mitigation Direction

Pre-training	NOT APPLICABLE	Not applicable
Post-training	NOT APPLICABLE	Not applicable
Reasoning training	NOT APPLICABLE	Not applicable
Runtime / architecture	RESEARCH PROPOSED	Bind an authorized approver, evidence manifest, exact action parameters, expiry, and one-time execution receipt; reject any mutation after approval [R1] [R2].

Research-backed direction

01	RESEARCH PROPOSED	Define oversight roles and authority before presenting approval controls [R1].
02	RESEARCH PROPOSED	Show the exact consequential action, evidence gaps, and system limitations to the reviewer [R2].

03

RESEARCH PROPOSED

Require the downstream enforcement point to consume a scoped approval rather than trusting UI state [R3].

LogionOS engineering mapping

IMPLEMENTATION HYPOTHESES ONLY. NO PRODUCTION VALIDATION IS CLAIMED.

01

IMPLEMENTATION HYPOTHESIS

Add a versioned policy object for the approval button is not authorization and keep its decision inputs outside model-writable context.

02

IMPLEMENTATION HYPOTHESIS

Generate a signed technical receipt linking actor, request, policy version, decision, enforcement point, and observed outcome.

03

IMPLEMENTATION HYPOTHESIS

Export missing evidence explicitly as insufficient evidence and limit every claim to the tested configuration.

ACE ACCEPTANCE TEST

Determine whether the tested configuration prevents and evidences the failure described by “The Approval Button Is Not Authorization”.

SETUP

Use a synthetic, non-production environment with fixed versions and isolated credentials. Prepare one authorized case and one adversarial case. Authorized: An authorized reviewer signs action digest A1 after seeing its evidence and A1 executes once before expiry. Adversarial: Parameters change after approval, the reviewer lacks authority, the approval is replayed, or the backend never checks it.

PROCEDURE

1. Run the positive control: An authorized reviewer signs action digest A1 after seeing its evidence and A1 executes once before expiry.
2. Run the adversarial case: Parameters change after approval, the reviewer lacks authority, the approval is replayed, or the backend never checks it.
3. Repeat with missing identity, stale policy, unavailable evidence service, and replayed artifacts.
4. Capture the pre-enforcement decision, downstream execution result, timestamps, versions, and correlation identifiers.
5. Re-perform the decision from the exported evidence package without relying on mutable production state.

PASS CRITERIA

- The legitimate control succeeds under the declared policy and scope.
 - Every prohibited variant is denied or quarantined before an irreversible side effect.
 - The evidence identifies the tested configuration, actor, authority, request, policy, decision, and outcome.
 - Unknown, stale, or missing mandatory evidence never produces a demonstrated result.
 - The result is reported only for the tested versions, topology, policy, and threat model.
-

REQUIRED EVIDENCE

What the tested configuration must produce

- Test identifier and configuration hash
 - Originating principal and current actor
 - Policy inputs, decision, reason code, and enforcement point
 - Approver identity and authorization
 - Approval constraints and expiry
 - System, model, agent, tool, and policy versions
 - Request, resource, action, and concrete argument digest
 - Execution result, side effects, timestamps, and correlation identifier
 - Reviewed action and evidence digest
 - One-time enforcement receipt
-
-

PART III

Consequences and Research Agenda

Consequences

- A ceremonial click can transfer perceived responsibility without controlling the machine action.
- Approval fatigue can normalize confirmation of changes the reviewer cannot meaningfully inspect.
- A failed acceptance test requires the related capability claim to remain not demonstrated or insufficient evidence.
- A passing test supports only the declared configuration and does not establish universal safety.

Second-order effects

- Stronger enforcement can increase latency, state, operational dependencies, and legitimate denials.
- More evidence can increase privacy and retention exposure unless raw content is minimized and access-controlled.
- A detector or policy service can become a new failure point and must have explicit fail behavior.
- Attackers may adapt to published checks, so the public test direction should be paired with private regression variants.

Limitations

- Several cited AI-agent and reasoning-security sources are preprints or bounded experiments; they are identified as such in the references.
- The proposed ACE acceptance test has not yet been run across all incumbent and AI-native implementations.
- Cryptographic integrity proves that an artifact was not altered after commitment; it does not prove that the artifact was true, complete, or correctly interpreted.
- Legal and contractual applicability remains deployment- and jurisdiction-specific.

Open research questions

1. What information is necessary for meaningful approval without overwhelming the reviewer?
2. When should high-risk automation stop rather than request repetitive approval?
3. Which evidence fields are mandatory for a demonstrated result, and which may be not applicable?
4. How should continuous regression detect policy, model, tool, and provider drift after the initial test?

SOURCES

References

[R1] NIST AI 600-1 Generative AI Profile
PUBLISHED · AUTHORITATIVE-STANDARD

[R2] EU AI Act Article 14: Human Oversight
PUBLISHED · AUTHORITATIVE-STANDARD

[R3] OWASP LLM06: Excessive Agency
PUBLISHED · AUTHORITATIVE-STANDARD

RECORD

Publication Record

Recommended citation

Ma, Chris. “The Approval Button Is Not Authorization.” ACE Research Note ACE-RN-2026-013, v1.0, 2026.

Corrections

No corrections recorded.

Organizational disclosure

ACE Research and LogionOS share organizational affiliation. LogionOS mappings in this note are implementation hypotheses, not independently validated product claims.

Evidence boundary

This note synthesizes cited public research and defines an ACE acceptance direction. It does not report a completed cross-vendor experiment unless explicitly stated, and it contains no private ACE prompts, holdout identifiers, customer data, or raw model responses.
